Machine Learning Engineer Nanodegree

Capstone Proposal

Dina Nashaat
January 14th, 2019

Proposal

Domain Background

In layman terms, Natural Language Processing (NLP) takes human language, as text or speech, as input and with some level of processing it provides useful contextual information in the field of translation, text analysis, classification, or data extraction. It combines interest and research in Linguistics, Computer Science, Artificial Intelligence and Logic in one field.

NLP has opened doors to various applications. To mention a few, It enables us now to read auto-generated text from speech, as in YouTube videos. Also, it has helped censor languages in public forums, automatically answering questions, converting text images to its corresponding text (OCR), deriving emotions from phrases, and topic segmentation.

Research in the field began in the 1950s (Jones, 2001), specifically tackling Machine Translation as a problem of research, one of the earliest encounters to the field was research done by IBM-Georgetown in Automatic translation from Russian to English (IBM, 1954). Since then, phases of research in NLP has transitioned from Machine Translation to Artificial Intelligence flavored applications, like the baseball question answering system (Bert F. Green, 1961), then focus diverted to statistical NLP, to provide naturalness, structure preference, and a degree of grammaticality to applications including speech recognition.

Problem Statement

The problem is currently held as a competition on Kaggle by Quora. Quora is an online platform where people post concise questions that can be answered by the community. A platform used by millions, and driven by the community is surely to suffer from insincere questions, those founded upon false premises, or that intend to make a statement rather than look for helpful answers, as described by Quora in their overview, in the competition page[1].

In this project, we will tackle this problem by developing a model that can flag questions as insincere to detect toxic and misleading content with more scalable methods.

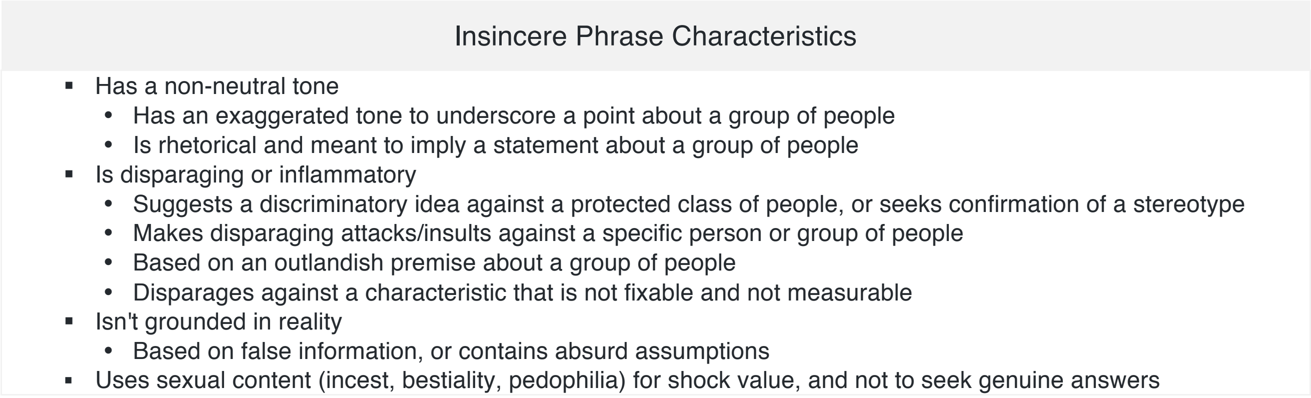
The problem is a classic NLP text classification problem, where a dataset of questions and posts are provided, and an insincere or not label, provided by manual review and machine learning.

As a supervised learning problem, it can be framed as quantifiable and measurable, using the pre-classified labels provided in the dataset. A lot of metrics provided by supervised learning can be employed to measure the score of the new model, which will be discussed below in Evaluation metrics.

Datasets and Inputs

In short, the problem is to detect whether a question asked on Quora is sincere or not. The training data is primarily composed of questions and a flag of 1 for insincere and 0 for otherwise.

The definition of an insincere question is given by Quora as: intended to make a statement rather than look for helpful answers, and they have defined a set of characteristics to flag a phrase as insincere, shown in Table 1.



Insincere Phrase Characteristics

- Has a non-neutral tone
  - Has an exaggerated tone to underscore a point about a group of people
  - Is rhetorical and meant to imply a statement about a group of people
- Is disparaging or inflammatory
  - Suggests a discriminatory idea against a protected class of people, or seeks confirmation of a stereotype
  - Makes disparaging attacks/insults against a specific person or group of people
  - Based on an outlandish premise about a group of people
  - Disparages against a characteristic that is not fixable and not measurable
- Isn't grounded in reality
  - Based on false information, or contains absurd assumptions
- Uses sexual content (incest, bestiality, pedophilia) for shock value, and not to seek genuine answers

Figure 1 Insincere phrase characteristics as defined by Quora's competition page on Kaggle

Two distinct datasets are provided by Quora as shown in Table 2, the first dataset is a training set with approximately 1.3 million entries, and 3 fields (qid, question text, target) as shown in Table 3. The second dataset is a test dataset with 2 fields (qid, question text), and the true classification for the test dataset is provided in the sample submission.

[1] https://www.kaggle.com/c/quora-insincere-questions-classification

Dataset		
Training Set	1.31 M x 3	A training set with 3 fields as shown in Table 2
Test set	56.4 K x 2	A test set with qid and question text
Sample Submission	56.4 K x 2	Submission file with qid and prediction

Figure 2 Datasets provided by Quora's competition page on Kaggle

Data Fields	
QID	Unique question identifier
Question Text	Quora question text
Target	Insincere has a value of 1 and 0 otherwise

Figure 3 : Data fields in the training set

Solution Statement

NLP text classification problem can be approached with different models, regular supervised algorithms, like Naïve Bayes or Support Vector Machine. However, for this particular problem I would like to propose a solution using Deep Learning, exploring CNN or RNN effects on this problem. CNN and RNN promise a wildly better accuracy over conventional supervised algorithms specifically Naïve Bayes as proposed by (Venkata Kishore Neppalli, 2018). Also, for this kind of problem, we would like to classify based on the semantic of the whole sentence/question, rather than relying on key words or small phrases.

The approach presented will lie heavily upon models provided by (Wenpeng Yiny, 2017), which compares CNN and RNN methods to multiple text based tasks, and CNN and RNN models provided by (Venkata Kishore Neppalli, 2018),

Benchmark Model

For this particular problem, I would like to provide two approaches for Benchmarking. The first approach is comparing metrics between the Naïve Bayes approach and the DNN (CNN or RNN) model presented with this project. The comparison will be in terms of performance (F1-Score).

The second approach is comparing how this model performs against the highest current F-Score (Jan. 14, 19) recorded on the competition leaderboard[2] which is 0.711.

[2] Quora Insincere questions classification leaderboard: https://www.kaggle.com/c/quora-insincere-questions-classification/leaderboard

Evaluation Metrics

For this kind of problem, I would propose using the F1-Score, which depends on the precision and recall.

$$F_1 Score = Harmonic\ Mean(Percision,\ Recall)$$

$$F_1 Score = 2\,\frac{Precision*Recall}{Precision+Recall}$$

Using the F1-Score will present a more suitable evaluation metric than accuracy because it takes false positives and false negatives into account, and it will always favor the less between Precision and Recall.

This also comes in advantage when comparing the performance for the second benchmarking approach, benchmarking against the competition leaderboard.

Project Design

Dataset Exploration

Explore a random subset from insincere and sincere questions, and identify any duplicate entries, provide visualization of each class proportion (bar chart).

Feature Engineering

First, we start by cleaning the textual data through a series of steps:

- Converting all records to lowercase
- Correcting typos and misspellings
- Removing punctuation
- Removing stop words
- Stemming and Lemmatization

To correct typos and misspellings, we will use the word2vec to approximate word probabilities, since word2vec orders words in decreasing order of frequency. Next, we will carry on removing punctuation, stop words, stemming and lemmatization using NLTK's library.

Tokenization

After cleaning the data, we start by tokenizing the input, among the different methods, we can use Keras Tokenizer API.

Building the Model

For this part, I would like to test first on a subset of the dataset, using a number of architectures, I would like to experiment using a CNN model and another RNN model (using LSTM) similar to a number of architectures in several literatures (Kim, 2014), (Venkata Kishore Neppalli, 2018), (Wenpeng Yiny, 2017) . However, adjusting layers and hyper-parameters will depend on the dataset exploration step. The winning model in terms of training time primarily, performance secondarily and prediction time, will be used to train the entire dataset, and will be presented as the final model.

Also, for the purpose of experimentation, I would like to consider applying transfer learning using word2vec and observe if it might provide a better prediction to the problem. Transfer learning might help fine tune the model since the training set is large, and has similarities with word2vec. However, this part of the model will be optional and only for experimentation, and will be considered if the previous models do not promise good performance.

Model Evaluation

Evaluate the performance of the chosen model by obtaining the F1 score.

References:

Bert F. Green, J. A. (1961). Baseball: An Automatic Question Answerer . *Proceedings of the Western Joint Computer Conference. 19*, pp. 219-224. Los Angeles, California : Western Joint Computer Conference.

IBM. (1954, January 8). *701 Translator: IBM Press Release*. Retrieved from IBM: https://www.ibm.com/ibm/history/exhibits/701/701_translator.html

Jones, K. S. (2001, October). Natural Language Processing: A Historical Review. *Artificial Intelligence Review* , 3-16.

Kim, Y. (2014). Convolutional Neural Networks for Sentence Classification. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*, (pp. 1746–1751.).

Venkata Kishore Neppalli, C. C. (2018). Deep Neural Networks versus Naïve Bayes Classifiers for Identifying Informative Tweets during Disasters. *ISCRAM Conference.* Rochester, NY, USA.

Wenpeng Yiny, K. K. (2017). Comparative Study of CNN and RNN for Natural Language Processing. *CoRR , abs/1702.01923*.